DOCUMENT RESUME

ED 168 849 SE 026 927

AUTHOR Webb, Norman L.; And Others
TITLE Mathematical Problem Solving Project Technical Report IV: Developmental Activities Related to Summative Evaluation (1975-1976). Final Report.
INSTITUTION Indiana Univ., Bloomington. Mathematics Education Development Center.
SPONS AGENCY National Science Foundation, Washington, D.C.
PUB DATE May 77
GRANT NSF-PES-74-15045
NOTE 37p.; For related documents, see SE 026 911-934; Contains occasional light and broken type

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Educational Research; Elementary Education; *Elementary School Mathematics; Evaluation; *Mathematics Education; *Problem Solving; *Summative Evaluation; *Tests
IDENTIFIERS *Mathematical Problem Solving Project; Research Reports

ABSTRACT

Four instruments were selected or developed for the summative evaluation of the Mathematical Problem Solving Project and pilot tested. The instruments were: (1) Student Attitude Questionnaire (SAQ), developed and validated by the MPSP evaluation staff; (2) problems selected from the National Longitudinal Study of Mathematics Achievement (NLSMA); (3) the problem-solving subtest of the Stanford Achievement Test (SAT); and (4) Problem Solving Survey (PSS), developed by the MPSP evaluation staff. A detailed report on the choice, development, and analysis of the instruments is included. (MP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

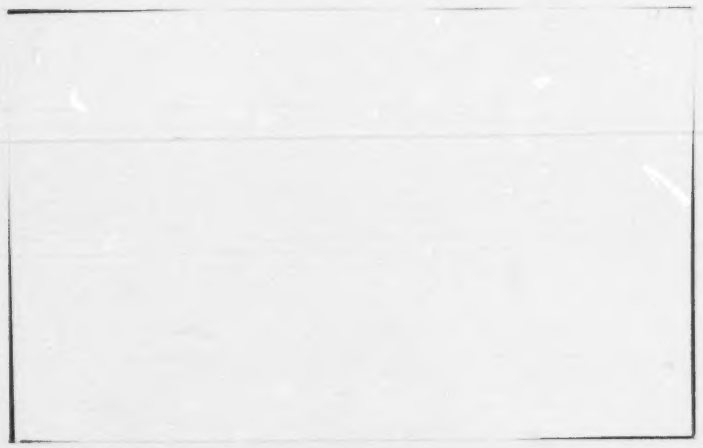
ED168849

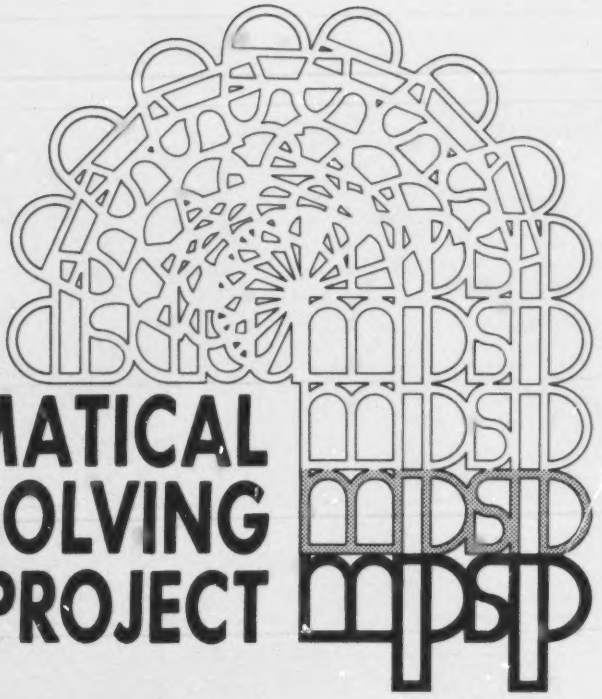
U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
John F. LeBlanc
George Immerzeel
David W. Wells

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."





MATHEMATICAL PROBLEM SOLVING PROJECT

SE 026 927

A Project of the
MATHEMATICS EDUCATION DEVELOPMENT CENTER
Project Supported by
National Science Foundation Grant PES74-15045

MATHEMATICAL PROBLEM SOLVING PROJECT

POLICY BOARD

George Immerzeel

Donald R. Kerr, Jr.*

John F. LeBlanc, *project director*

George Springer, *co-principal investigator*

Maynard Thompson

David W. Wells*

* *acting co-directors, 1975-76*

ADVISORY BOARD

Robert P. Dilworth,
California Institute of Technology

James F. Gray,
St. Mary's University

John L. Kelley,
University of California at Berkeley

Jeremy Kilpatrick,
Teachers College, Columbia University

Eugene D. Nichols,
Florida State University

UNIVERSITY OF NORTHERN IOWA CENTER
Cedar Falls, Iowa

George Immerzeel,
center director
Joan E. Duea
Earl G. Ockenga
John E. Tarr
Jack D. Wilkinson

INDIANA UNIVERSITY CENTER
Bloomington, Indiana

John F. LeBlanc,
center director
Marilyn Hall Jacobson
Donald R. Kerr, Jr.**
Frank K. Lester, Jr.
George Springer
Arthur H. Stengel
Maynard Thompson
Norman L. Webb

OAKLAND SCHOOLS CENTER
Pontiac, Michigan

David W. Wells,
center director
Stuart A. Choate
Douglas W. MacPherson

***acting center director, 1975-76*

Project Graduate Students: Randall I. Charles, Fadia F. Harik, Tom S. Hudson, Barbara E. Moses, Gary Post, Linda A. Proudfit

FINAL REPORT
MATHEMATICAL PROBLEM SOLVING PROJECT
TECHNICAL REPORT IV:
DEVELOPMENTAL ACTIVITIES RELATED TO SUMMATIVE EVALUATION (1975-1976)

Report Prepared By

Norman L. Webb
Project Evaluator

Barbara E. Moses
Evaluation Assistant

Donald R. Kerr, Jr.
Evaluation Director
Project Assistant Director

Under the Direction of
John F. LeBlanc
Project Director

MATHEMATICS EDUCATION DEVELOPMENT CENTER
Indiana University - Bloomington
May 1977

TABLE OF CONTENTS

		Page
A.	*RATIONALE FOR SUMMATIVE EVALUATION*	1
B.	*INSTRUMENTS AND PROCEDURES*	2
C.	*ATTITUDE MEASUREMENT*	3
	1. Rationale	3
	2. The SAQ	3
	3. SAQ Data	4
	4. Analysis of SAQ Data	4
D.	*NLSMA SUBSCALE ITEMS*	8
	1. Rationale	8
	2. NLSMA Data	8
	3. Analysis of NLSMA Data	12
E.	*STANFORD ACHIEVEMENT TEST*	12
	1. Rationale	12
	2. SAT Data	13
F.	*PROBLEM SOLVING SURVEY*	13
	1. Rationale	13
	2. PSS Data	15
G.	*SUMMARY*	15
	1. SAQ	15
	2. NLSMA	19
	3. SAT	19
	4. PSS	19
	5. MPSP	19
MPSP WORKING PAPER 1975-76: FORMATION OF TESTS FOR THE SUMMATIVE EVALUATION		21

A. *Rationale for Summative Evaluation*

The Mathematical Problem Solving Project (MPSP) had a long-range intention of impacting the materials and procedures of teaching mathematics to children. So it was clear from the beginning that there would be a time when the total impact of MPSP materials and procedures on children would have to be measured.

The project goal was to explore ways to improve the problem-solving abilities of elementary school children. More specifically, the project wanted to foster:

(1) improved achievement on nonstandard, process-type problems;

(2) the use of a richer repertoire of skills and strategies in problem solving; and as a spinoff of (1) and (2),

(3) improved achievement on standard textbook problems.

To measure the attainment of specific goals, it is important to have instruments that are familiar and accepted. It is also important to have instruments which are sensitive to these specific goals. During the 1975-76 school year, the goals of the MPSP summative evaluation were to identify existing instruments to measure its goals, to develop instruments where no satisfactory ones existed, and to pilot-test the instruments and procedures for administering them. This is a report of the selection, development, and pilot-testing of instruments used in the 1975-76 MPSP summative evaluation.

Since MPSP was terminated, it has become clear that these efforts will represent the only systematic evaluation of the project. So this report will cover both instrument selection and testing and any tentative observations concerning the project that can be based on the data from the pilot-testing of these instruments. It is important to note that the treatment received by the children involved in this study was

not the systematic application of a developed program but rather experiences with preliminary materials which were under development. While every effort was made to insure that the children had experiences that were educationally sound, these experiences were in no way based on a completely developed program.

B. *Instruments and Procedures*

This report is organized around the four instruments selected or developed for the summative evaluation and used during the 1975-76 school year. The instruments used for the summative evaluation were:

1. Student Attitude Questionnaire (SAQ), developed and validated by the MPSP evaluation staff;
2. problems selected from the National Longitudinal Study of Mathematics Achievement (NLSMA);
3. the problem-solving subtest of the Stanford Achievement Test (SAT);
4. Problem Solving Survey (PSS), developed by the MPSP evaluation staff.

Each of these instruments is briefly discussed below with respect to the procedures used for developing the instrument or the process used for selecting the instrument, how the instrument was used, the data collected from the instrument, the implications of the data, and what was learned about the instrument. A detailed report on the choice, development, and analysis of the instruments is included in the Working Paper which begins on page 21.

The children involved in the testing reported here were those students in the Oakland Schools that received all three MPSP problem-solving modules during the 1975-76 school year. In addition to the experimental classes (i.e., the classes that received the problem-solving modules), control classes

were identified.* Both the experimental and control classes were given each of the four instruments in the fall 1975, before the experimental classes used any of the modules. Each of the four instruments was administered again to all experimental and control classes in the spring 1976. All instruments were administered by the classroom teachers.

C. *Attitude Measurement*

1. Rationale: A year of exploration with children and a review of the problem-solving literature suggested that willingness, confidence, and perseverance are three factors that influence problem-solving performance. Consequently, the MPSP staff felt it was important to measure the impact of MPSP on students' willingness to solve problems, confidence in their ability to solve problems, and perseverance in attempting to obtain a solution.

The history of paper-and-pencil attitude measurement is marked with few successes. Most attitude studies result in no significant differences; and even when there are significant differences, they often prove difficult to interpret. MPSP was looking for treatment-specific attitude change (i.e., specific to problem solving) rather than changes in general attitudes. Since the literature on attitude testing does contain some successes with treatment-specific instruments, MPSP felt there was some hope of identifying treatment-specific attitude changes.

2. Development of the SAQ: The details of the development and pilot analysis of the SAQ are contained on pp. 23-27 of the Working Paper and in Appendices B, C, D, E, and F. The SAQ is a self-report, paper-and-pencil instrument which contains subscales intended to measure

* See Technical Report III, Chapter A for a detailed description of the experimental and control classes.

willingness, confidence and perseverance. The pilot analysis of the SAQ suggested that the SAQ measures some aspects of the affect involved in mathematical problem solving and that the three subscales were distinct and had moderate internal consistency.

3. SAQ Data: The data obtained for the SAQ is reported in Tables 1, 2 and 3 for the 4th, 5th and 6th grades. In reporting this data, the classes were separated into thirds (lower, middle and upper) based on their scores on the SAT. Scores for the lower third are reported after "Low," for the upper third after "High," and for all students after "Whole Class." Pretest, post-test, and gain scores are reported for both experimental and control groups for each of the three subscales (willingness, perseverance and self-confidence) as well as for the whole test.

At the fourth grade, trends favor the experimental group with a smattering of significant differences ($p<.05$). At the fifth grade those trends were reversed. At the sixth grade, results were different for different subscales with a slight overall edge to the experimental group. Looking at high-ability children and low-ability children, one notes that high-ability students scored consistently higher than low-ability students.

4. Analysis of SAQ Data: There is no clear trend in the pretest and post-test comparisons for the experimental and control groups. The pilot efforts with the SAQ indicated that some components of student attitude are measured by this instrument. However, the data suggests that the MPSP experience did not have a consistent impact on the components of student attitude measured by the SAQ. All of the teachers in the experimental classes reported positive student attitudes toward problem solving related to MPSP, yet no consistent trends favoring the

Table 1
Comparison of the Experimental and Control
Fourth-Grade Classes on the SAQ

	Experimental			Control			
	Pre[a] (N=130)	Post (N=124)	Gain[b] (N=121)	Pre (N=65)	Post (N=73)	Gain (N=61)	F[c]
Willingness (max.=6)							
Low:[d]	3.600	3.917	0.375	4.529	4.333	-0.036	0.480
High:	4.961	4.673	-0.234	5.000	5.000	-0.083	0.017
Whole Class:	4.213	4.281	0.068	4.631	4.519	-0.112	0.542
Perseverance (max.=5)							
Low:	3.953	4.270	0.286	3.784	3.830	-0.036	1.751
High:	4.189	4.333	0.188	4.308	4.538	0.231	0.527
Whole Class:	4.106	4.203	0.097	3.986	4.063	0.077	0.468
Self-Confidence (max.=7)							
Low:	3.875	4.194	0.281	4.189	3.619	-0.536	6.437**
High:	5.385	5.653	0.313	5.385	5.154	-0.231	2.765
Whole Class:	4.723	4.906	0.183	4.391	4.095	-0.296	6.157
SAQ Total (max.=18)							
Low:	11.649	12.400	0.679	12.647	11.951	-0.560	2.090
High:	14.480	14.646	0.333	14.583	14.692	-0.167	0.322
Whole Class:	13.169	13.452	0.283	13.062	12.795	-0.267	1.337

** $p<.05$

[a] The N of students represents the number of students used to compute the whole group mean. The students used for computing gain scores were the students that completed both the pre- and posttests.

[b] Gain = posttest score - pretest score.

[c] ANCOVA was used to test the difference in gain score means. The pretest for each student was used as the covariate.

[d] Low and high were determined by the lower and upper thirds of the distribution of SAT scores for the fourth-grade students.

Table 2
Comparison of the Experimental and Control Fifth-Grade Classes on the SAQ

	Experimental			Control			
	Pre[a] (N=202)	Post (N=188)	Gain[b] (N=181)	Pre (N=75)	Post (N=74)	Gain (N=70)	F[c]
Willingness (max.=6)							
Low:[d]	3.769	4.136	0.301	4.100	4.364	0.462	0.381
High:	4.803	4.864	0.169	4.545	4.714	0.000	0.556
Whole Class:	4.341	4.492	0.151	4.359	4.513	0.154	0.003
Perseverance (max.=5)							
Low:	3.846	3.608	-0.338	3.767	3.818	0.269	0.758
High:	4.212	4.117	-0.051	4.045	4.318	0.273	3.831*
Whole Class:	3.991	3.918	-0.073	3.974	4.013	0.039	1.609
Self-Confidence (max.=7)							
Low:	3.829	3.618	-0,308	3.645	3.484	0.038	0.069
High:	5.485	5.533	0.169	5.591	5.810	0.000	0.057
Whole Class:	4.675	4.560	-0.115	4.449	4.513	0.064	0.392
SAQ Total (max.=18)							
Low:	11.527	11.521	-0.323	11.379	11.710	1.000	1.058
High:	14.609	14.627	0.351	14.182	15.100	0.250	0.080
Whole Class:	13.104	13.094	-0.010	12.880	13.176	0.296	0.160

* $p<.10$

[a] The N of students represents the number of students used to compute the whole group mean. The students used for computing gain scores were the students that completed both the pre- and posttests.

[b] Gain = posttest score - pretest score.

[c] ANCOVA was used to test the difference in gain score means. The pretest for each student was used as the covariate.

[d] Low and high were determined by the lower and upper thirds of the distribution of SAT scores for the fifth-grade students.

Table 3
Comparison of the Experimental and Control
Sixth-Grade Classes on the SAQ

	Experimental			Control			
	Pre[a] (N=280)	Post (N=270)	Gain[b] (N=262)	Pre (N=84)	Post (N=83)	Gain (N=83)	F[c]
Willingness (max.=6)							
Low:[d]	4.152	4.038	-0.143	4.590	4.000	-0.526	0.111
High:	4.800	4.755	0.000	5.100	5.143	0.000	0.977
Whole Class:	4.329	4.349	0.020	4.741	4.430	-0.311	0.244
Perseverance (max.=5)							
Low:	3.626	3.739	-0.115	4.128	3.810	-0.297	0.034
High:	4.130	4.021	-0.085	4.300	4.214	-0.071	0.440
Whole Class:	3.881	3.833	-0.048	4.129	3.849	-0.280	0.057
Self-Confidence (max.=7)							
Low:	4.154	4.104	-0.034	3.900	3.675	-0.167	1.366
High:	4.673	5.695	0.109	5.733	5.964	0.250	0.696
Whole Class:	4.825	4.857	0.032	4.721	4.627	-0.094	1.376
SAQ Total (max.=18)							
Low:	12.173	12.091	-0.321	12.711	11.700	-0.857	0.828
High:	14.633	14.447	-0.022	15.133	15.321	0.179	0.934
Whole Class:	13.164	13.204	0.040	13.655	13.072	-0.583	0.957

[a] The N of students represents the number of students used to compute the whole group mean. The students used for computing gain scores were the students that completed both the pre- and posttests.

[b] Gain = posttest score - pretest score.

[c] ANCOVA was used to test the difference in gain score means. The pretest for each student was used as the covariate.

[d] Low and high were determined by the lower and upper thirds of the distribution of SAT scores for the sixth-grade students.

experimental students were found in the SAQ data. This inconsistency between the teachers' feedback and the SAQ data is difficult to interpret. The teachers may have been "judging" students' attitudes on components other than those measured by the SAQ. Another possible explanation is that the SAQ may be too general an instrument to measure attitudes toward solving problems of the type primarily used by MPSP. Whatever the reason for the inconsistency, it is clear that further efforts are warranted to refine or replace the SAQ as an instrument for measuring attitudes toward problem solving as defined by MPSP.

D. *NLSMA Subscale Items*

1. Rationale: The MPSP staff wanted to identify an instrument that was well-known to the mathematics education community and which measured problem-solving performance on the kinds of problems emphasized by MPSP. An extensive review of the literature did not identify any such instrument. However, among the problems used in the National Longitudinal Study for Mathematics Achievement (NLSMA) several items in one subscale were similar to the process-type problems used by MPSP, and the NLSMA tests are nationally known and widely accepted. Three items (1-3) were chosen from NLSMA more as traditional, textbook-type problems which would not be threatening to children on a pretest. Two items (4 and 5) were selected as process-type problems. The details of the selection process of the five-item, multiple choice NLSMA subscale appear in the discussion of the Problem Solving Survey (part I) on page 28 of the Working Paper (see the NLSMA items in Appendix G).

2. NLSMA Data: Tables 4, 5 and 6 show the data obtained from the NLSMA. For fourth and sixth grades, the gains consistently favored experimental students over control. Many of the differences were at or near significance ($p<.05$). This picture is even clearer for items 4 and 5, which are most like those problems emphasized by MPSP.

Table 4
Comparison of the Experimental and Control
Fourth-Grade Classes on NLSMA

Items	Experimental			Control			F[c]
	Pre[a] (N=123)	Post (N=121)	Gain[b] (N=121)	Pre (N=59)	Post (N=76)	Gain (N=57)	
1,2,3 (max.=3)							
Low:[d]	1.395	1.139	-0.226	0.886	0.957	0.000	0.021
High:	1.569	1.898	0.292	1.615	1.833	0.083	3.862*
Whole Group:	1.489	1.558	0.069	1.200	1.165	-0.035	7.188***
4,5 (max.=2)							
Low:	0.351	0.788	0.370	0.382	0.523	0.143	2.661
High:	0.813	1.250	0.419	0.800	1.000	0.100	2.057
Whole Group:	0.579	1.016	0.437	0.516	0.688	0.172	5.324**
All (max.=5)							
Low:	1.750	1.938	0.231	1.219	1.523	0.111	3.014
High:	2.391	3.146	0.667	2.700	2.750	0.100	0.094
Whole Group:	2.073	2.603	0.530	1.729	1.868	0.139	7.761***

*** $p<.01$

** $p<.05$

* $p<.10$

[a] The N of students represents the number of students used to compute the whole group mean. The students used for computing gain scores were the students that completed both the pre- and posttests.

[b] Gain = posttest score - pretest score.

[c] ANCOVA was used to test the difference in gain score means. The pretest for each student was used as the covariate.

[d] Low and high were determined by the lower and upper thirds of the distribution of SAT scores for the fourth-grade classes.

Table 5
Comparison of the Experimental and Control
Fifth-Grade Classes on NLSMA

Items	Experimental			Control			F^c
	Pre[a] (N=190)	Post (N=188)	Gain[b] (N=181)	Pre (N=71)	Post (N=77)	Gain (N=70)	
1,2,3 (max.=3)							
Low:[d]	1.026	1.289	0.309	0.966	1.344	0.160	0.975
High:	2.045	2.200	0.217	1.667	2.182	0.476	0.049
Whole Group:	1.466	1.686	0.220	1.171	1.747	0.576	0.570
4,5 (max.=2)							
Low:	0.373	0.625	0.310	0.333	0.250	-0.174	13.148***
High:	0.841	1.153	0.386	0.900	1.182	0.400	0.126
Whole Group:	0.568	0.894	0.326	0.611	0.615	0.004	9.951***
All (max.=5)							
Low:	1.388	1.903	0.638	1.333	1.581	-0.087	7.334***
High:	2.889	3.373	0.614	2.526	3.364	0.895	0.065
Whole Group:	2.047	2.585	0.538	1.789	2.351	0.562	2.088

*** $p<.01$

[a] The N of students represents the number of students used to compute the whole group mean. The students used for computing gain scores were the students that completed both the pre- and posttests.

[b] Gain = posttest score - pretest score.

[c] ANCOVA was used to test the difference in gain score means. The pretest for each student was used as the covariate.

[d] Low and high were determined by the lower and upper thirds of the distribution of SAT scores for the fifth-grade students.

Table 6
Comparison of the Experimental and Control
Sixth-Grade Classes on NLSMA

Items	Experimental			Control			F[c]
	Pre[a] (N=276)	Post (N=281)	Gain[b] (N=275)	Pre (N=84)	Post (N=84)	Gain (N=84)	
1,2,3 (max.=3)							
Low:[d]	1.324	1.709	0.384	1.200	1.512	0.297	0.555
High:	2.287	2.484	0.189	2.400	2.179	-0.214	0.649
Whole Group:	1.837	2.070	0.233	1.744	1.835	0.091	3.553*
4,5 (max.=2)							
Low:	0.469	0.824	0.395	0.474	0.619	0.111	4.024**
High:	0.794	1.319	0.522	0.900	1.074	0.148	3.619*
Whole Group:	0.615	1.089	0.474	0.631	0.800	0.169	12.416***
All (max.=5)							
Low:	1.811	2.533	0.824	1.658	2.122	0.486	3.144*
High:	3.082	3.819	0.711	3.300	3.259	-0.037	8.275***
Whole Group:	2.489	3.141	0.652	2.381	2.631	0.250	11.958***

*** $p<.01$

** $p<.05$

* $p<.10$

[a] The N of students represents the number of students used to compute the whole group mean. The students used for computing gain scores were the students that completed both the pre- and posttests.

[b] Gain = posttest score - pretest score.

[c] ANCOVA was used to test the difference in gain score means. The pretest for each student was used as the covariate.

[d] Low and high were determined by the lower and upper thirds of the distribution of SAT scores for the sixth-grade students.

The results for the fifth grade are less clear. On the whole test there is a slight trend toward the low-ability experimental students, and there is a trend toward the high-ability control students. However, there is a significant edge to experimental students on problems 4 and 5. It should be noted that conflicting and sometimes surprising results for the fifth-grade students have persisted throughout both the formative and summative evaluations.

3. Analysis of NLSMA Data: The NLSMA subscale proved to be easy to administer and easy to score. Problems 1, 2 and 3 did provide a sufficient amount of success to avoid any difficulties in using the test as a pretest or with control students. The entire NLSMA test has national acceptance, and the subscale (especially items 4 and 5) has face validity with respect to MPSP goals. The data collected suggests that the NLSMA problems (especially items 4 and 5) are sensitive to the MPSP treatment. The set of problems seems appropriate for use in any future MPSP evaluations.

E. *Stanford Achievement Test* (SAT)

1. Rationale: The use of a standardized achievement test was decided upon for two reasons. First, the project hoped to demonstrate that on a widely accepted instrument, the MPSP treatment did not result in a decline in achievement in traditionally-measured areas of mathematics. Secondly, the project was interested in knowing if the MPSP treatment had sufficient impact so that it could be detected on a kind of test that has traditionally proved to be insensitive to treatment effects. By design, standardized achievement tests measure those skills which are basic to all school curricula. Although many research and evaluation studies have reported data from these tests, few have found significant differences in means due to treatment.

In choosing an appropriate achievement test, one of the primary criteria was that the problem-solving portion have as much emphasis as possible on process-type problems. The Stanford Achievement Test was chosen because the problem-solving portion had a spirit that was most similar to that of MPSP and because it ranked at least as high as the others on other criteria (e.g., readability). The details of the achievement test selection are reported in Appendix A. The SAT itself is Appendix H.

2. SAT Data: In order to limit the total testing time to an acceptable level, only the problem-solving portion of SAT Intermediate Level I was given. It was administered to experimental and control students on a pretest - posttest basis.

The data shows consistent trends in favor of the experimental classes. While the data does not contain consistent significant differences, it clearly supports the claim that MPSP experiences do not result in losses in more traditional problem-solving skills. (See Table 7.)

The SAT was easily administered and seemed to exhibit some sensitivity to the MPSP treatment. If testing time had allowed, it would have been good to administer the computation portion of the test to support a wider claim concerning the lack of negative side effects (namely, a decrease in computational skills) due to MPSP.

F. *Problem Solving Survey (PSS)*

1. Rationale: The project wanted an instrument that was sensitive to the MPSP goals, that would provide some success experiences for most children, and that would provide some insight into the type of problem-solving processes that were actually being used by children. To meet these needs the PSS was developed. It was a three-problem, open-ended test (i.e., not multiple choice). The first problem on each form (four

Table 7
Comparison of the Experimental and Control SAT Means

Grade		Experimental			Control			F[b]
		Pre	Post	Gain[a]	Pre	Post	Gain	
4	No. of students	133	116	115	66	79	64	
	Mean	27.564[c]	30.767	3.203	22.894	24.291	1.397	3.119*
5	No. of students	207	185	181	68	71	63	
	Mean	27.681	29.605	1.924	27.971	29.577	1.606	0.233
6	No. of students	286	266	262	72	82	70	
	Mean	32.500	34.598	2.098	32.028	32.841	0.813	2.392

* $p<.10$

[a] Gain = postest score - pretest score.

[b] ANCOVA was used to test the differences in group means. The pretest score for each student was used as the covariate.

[c] Maximum was 40.

forms for each of three grade levels) was a simple, one-step, textbook-type word problem. The second problem was a multiple-step textbook-type problem, and the third problem was a process-type problem of the kind focused on by MPSP. The student was encouraged to show all of his/her work on the paper and was given an example of what this meant in the instrument's instructions. The details of the development of the PSS are contained in the Working Paper on pages 28-30. (The PSS is referred to there as Part II of the Problem Solving Survey. The PSS, itself, is Appendix I.)

2. PSS Data: The PSS was administered in the same way as the other instruments and the results were analyzed in two ways. First, the tests were evaluated on a "right-wrong" basis. The data from this evaluation was analyzed and reported in the same way as the data for the other instruments (see Tables 8, 9 and 10).

For the correct-answer analysis, the fourth-grade experimental students had a significantly higher mean gain score than the fourth-grade control students ($p<.01$). At the fifth- and sixth-grades, the mean gain scores for the experimental students were higher than the mean gain scores for the control students but the differences were not significant. On the whole, the instrument did seem to have some sensitivity to MPSP goals, and problem (1) did seem to provide the desired success experience for the students as evidenced by the high group mean for this problem.

A second evaluation of the test was an informal comparison of the pre- and posttest papers to see if the PSS had any potential for reflecting student problem-solving processes. Growth in experimental students was observed with respect to the amount of work shown, as well as with respect to the use of tables to solve problems. The conclusion was that

Table 8
Comparison of Experimental and Control
Fourth Grade-Classes on the PSS

	Experimental			Control			
	Pre[a] (N=126)	Post (N=119)	Gain[b] (N=112)	Pre (N=64)	Post (N=55)	Gain (N=52)	F[c]
Problem 1							
Low:[d]	0.568	0.706	0.121	0.500	0.556	0.077	.056
High:	0.800	0.848	0.043	0.917	0.917	0.083	.178
Whole Group:	0.698	0.756	0.058	0.609	0.492	-0.117	2.867*
Problem 2							
Low:	0.108	0.324	0.242	0.056	0.148	0.115	2.147
High:	0.560	0.565	-0.043	0?	0.417	0.149	.537
Whole Group:	0.365	0.437	0.072	0.094?	0.218	0.124	5.026**
Problem 3							
Low:	0.222	0.388	0.212	0.161	0.089?	0.038	7.261***
High:	0.476	0.591	0.096	0.350	0.467	0.117	.156
Whole Group:	0.367	0.501	0.134	0.219	0.273	0.054	3.610*
Total							
Low:	0.897	1.418	0.576	0.717	0.793	0.231	4.486**
High:	1.836	2.004	0.096	1.267	1.800	0.533	1.769
Whole Group:	1.430	1.694	.264	.922	1.055	.133	8.711***

*** $p<.01$
** $p<.05$
* $p<.10$

[a] The N of students represents the number of students used to compute the whole group mean. The students used for computing gain scores were the students that completed both the pre- and posttests.

[b] Gain = posttest score - pretest score.

[c] ANCOVA was used to test the difference in gain score means. The pretest for each student was used as the covariate.

[d] Low and high were determined by the lower and upper thirds of the distribution of SAT scores for the fourth-grade students.

Table 9
Comparison of Experimental and Control
Fifth-Grade Classes on the PSS

	Experimental			Control			
	Pre[a] (N=187)	Post (N=168)	Gain[b] (N=162)	Pre (N=68)	Post (N=63)	Gain (N=61)	F[c]
Problem 1							
Low:[d]	0.417	0.682	0.246	0.484	0.654	0.208	.014
High:	0.820	0.891	0.075	0.778	0.889	0.111	1.103
Whole Group:	0.620	0.786	0.166	0.603	0.825	0.222	.813
Problem 2							
Low:	0.097	0.197	0.077	0.161	0.231	0.125	.143
High:	0.443	0.764	0.340	0.333	0.556	0.222	2.254
Whole Group:	0.283	0.446	0.163	0.206	0.381	0.175	.368
Problem 3							
Low:	0.091	0.247	0.147	0.055	0.130	0.078	1.222
High:	0.332	0.515	0.192	0.339	0.429	0.090	.685
Whole Group:	0.201	0.393	0.192	0.161	0.255	0.094	4.236**
Total							
Low:	0.605	1.126	0.470	0.700	1.014	0.411	.016
High:	1.594	2.170	0.607	1.450	1.874	0.424	.907
Whole Group:	1.105	1.625	0.520	0.969	1.461	0.492	.635

** $p<.05$

[a] The N of students represents the number of students used to compute the whole group mean. The students used for computing gain scores were the students that completed both the pre- and posttests.

[b] Gain = posttest score - pretest score.

[c] ANCOVA was used to test the difference in gain score means. The pretest for each student was used as the covariate.

[d] Low and high were determined by the lower and upper thirds of the distribution of SAT scores for the fifth-grade students.

Table 10
Comparison of the Experimental and Control
Sixth-Grade Classes on the PSS

	Experimental			Control			
	Pre[a] (N=285)	Post (N=277)	Gain[b] (N=270)	Pre (N=73)	Post (N=69)	Gain (N=68)	F[c]
Problem 1							
Low:[d]	0.547	0.743	0.196	0.529	0.774	0.226	1.972
High:	0.787	0.849	0.065	0.750	0.826	0.087	.170
Whole Group:	0.667	0.787	0.120	0.616	0.754	0.138	.107
Problem 2							
Low:	0.170	0.267	0.076	0.088	0.161	0.097	.410
High:	0.447	0.570	0.130	0.458	0.348	-0.130	4.702**
Whole Group:	0.284	0.426	0.142	0.233	0.232	0.001	7.511***
Problem 3							
Low:	0.268	0.446	0.173	0.294	0.323	0.032	.686
High:	0.493	0.585	0.107	0.422	0.658	0.217	.789
Whole Group:	0.389	0.519	0.130	0.324	0.538	0.214	.155
Total							
Low:	0.985	1.455	0.444	0.912	1.258	0.355	.041
High:	1.727	2.004	0.303	1.630	1.832	0.174	.738
Whole Group:	1.339	1.732	0.393	1.173	1.524	0.351	1.132

*** $p<.01$
** $p<.05$

[a] The N of students represents the number of students used to compute the whole group mean. The students used for computing gain scores were the students that completed both the pre- and posttests.

[b] Gain = posttest score - pretest score.

[c] ANCOVA was used to test the difference in gain score means. The pretest for each student was used as the covariate.

[d] Low and high were determined by the lower and upper thirds of the distribution of SAT scores for the sixth-grade students.

a careful analysis of PSS papers could reveal group changes in the use of problem-solving processes, especially in those processes which are likely to be manifested on paper (e.g., making a table, using a list, drawing a picture, writing an equation, etc.). It was decided that in order to make such an analysis meaningful, the PSS would have to be considerably refined, especially as to the selection of problems of type 3.

G. *Summary*

1. SAQ: It is clear that in order to measure attitude change due to the MPSP treatment, more work is needed. Informal feedback from teachers and children suggests that the MPSP treatment had a positive impact on student attitudes. It seems that in order to verify this impact, either the informal feedback procedures will have to be formalized (e.g., formal interviews and questionnaires) or considerably more refinement will have to take place in the SAQ.

2. NLSMA: The NLSMA subscale seems to have been a consistent and sensitive instrument and should probably be considered in any future evaluation efforts.

3. SAT: The SAT has limited potential for reflecting growth due to the MPSP treatment. However, it would be good to administer the entire SAT to refute possible claims that the MPSP treatment results in decreases in traditional mathematical skills.

4. PSS: The PSS seems to provide a non-threatening and sensitive approach to measuring MPSP goals. If the process evaluation use of PSS is to be pursued, additional attention needs to be given to determining the processes likely to be used on each problem and to developing reliable rating procedures.

5. Implications for MPSP: Given the fact that the intent of the MPSP treatments during 1975-1976 was to provide formative trials for the problem-solving modules and not to provide a complete and coherent program for the children, and given the fact that the purpose of the summative evaluation effort was to pilot-test instruments and procedures and not to evaluate MPSP, the outcomes were fairly encouraging for the project. Trends in the data generally favored the experimental students. Also, some instances were noted where the mean gain score for the experimental students was significantly higher than the mean gain score for the control students on process-type problems. On more traditional types of problems, the experimental students generally performed as well as the control students and in some instances outgained the control students. The summative evaluation results related to student attitudes are not clear.

MPSP Working Paper 1975/76

Formation of Tests for the Summative Evaluation

Norman L. Webb

The evaluation efforts for the school year 1975/1976 are divided into two parts, formative and summative. The formative evaluation will focus on each module with the goal of improving the effectiveness of the modules in attaining their objectives. The summative evaluation will focus on the assessment of changes in the students and teachers over the year. A more detailed explanation of the evaluation design is given in A Description of an Evaluation Design for the Mathematical Problem Solving Project: Rationale and Proposed Implementation (Lester and Webb, 1975). The purpose of this working paper is to delineate the rationale and procedures used to compile and create the tests for the pre- posttesting for the summative evaluation.

At the June 9, 10, and June 23, 24, 1975 MPSP staff meetings it was agreed by representatives of all three centers to include the following instrumentation in the test battery to be used for the summative evaluation:

1) A "problem-solving" portion of a standardized test that was available for administration by the Oakland Center.
2) A paper-and-pencil instrument to assess the child's preference of problems and confidence in his ability to solve problems.
3) A problem-solving survey composed of problems at different cognitive levels, including problems of the type that the project considers important.
4) A problem-sort task that will be given to the teachers to assess their perception of the appropriateness of problems for

use in teaching problem solving and their perception of problems which would be preferred by children. (Note: This sort task was agreed upon at the June 9, 10 meeting and was administered to the project teachers on June 17.)

5) An interview and observation format that teachers can use to evaluate their students' willingness and perseverance with respect to problem solving.

The dates set for the administration of the tests to the students were early October and May. The dates for administering the sort-task to the teachers were June 1975 and June 1976 at the first and last in-service meeting of the teachers for the year. It was also agreed upon to give the testing to a control group of approximately ten classes.

Standardized test

The Oakland Center sent to I.U. copies of all the standardized tests they had in-house and available for administering. The five tests included were: Iowa Tests of Basic Skills (Forms 5 & 6), Stanford Achievement Tests (Intermediate Levels I and II Form 4), Metropolitan Achievement Tests (Intermediate Form F and Elementary Form F), California Achievement Tests, and the SCAT-STEP (Series II, Form FA). These tests were reviewed systematically and judged on their capability to measure problem-solving ability. Very few differences were found among tests. Test 6, Math Application, of the Stanford Achievement Test (Intermediate Level I) was felt more appropriate for our needs. The selection process is reported in "Evaluation of Standardized Tests" (Charles and Moses, 1975) (see Appendix A). A copy of Test 6 is included in Appendix H.

Student Attitude Questionnaire (SAQ)

The SAQ is a paper-pencil questionnaire designed to measure different affective aspects of the problem-solving processes of students. The SAQ has gone through several changes during its evolution.

Initially the SAQ had three parts. Part I contained fourteen statements about solving word problems. The student was asked to indicate whether he agreed or disagreed with each statement. The items were divided into three scales: willingness to engage in problem-solving activities (4 items), perseverance during the problem-solving processes (4 items), and self-confidence with respect to problem solving (6 items). These items were selected from a list of 136 items collected from several sources including the Childhood Attitude Inventory for Problem Solving developed by Covington (1966) for use with the Productive Thinking Program. I.U. Staff members (total of 10) individually sorted the 136 items into four categories. Three of the categories are mentioned above. The fourth was curiosity about the solving of a problem. Within each category the raters ranked the items according to how well the item expressed the attitude. Items in the curiosity category did not have any cohesiveness or the appearance of forming one attribute of the affective domain. This category was dropped.

Items that all could agree upon to fit in a category and ranked high within that category by several people were used in Part I of the SAQ. Care was taken to ensure that within one category positive and negative items were evenly represented.

Part II was designed to have students respond to different types of problems with respect to the three categories of willingness, perseverance,

and self-confidence. Four problems were included. Two were word problems like those found in textbooks, one was a process-puzzle type problem (a problem from the project's problem deck), and one was a group-project problem. Three statements, reflecting each of the three categories, were listed after each problem. The students were to respond whether they agreed or disagreed with the statement with respect to the given problem.

Part III was designed to assess the student's preference as to the type of problems he/she would solve. Two sets of three problems were given. Each set represented a range of problem types (textbook, process-puzzles, and group-project). For each set of problems, the students were asked to select the problem they would most like to solve and then the problem they would next like to solve if given a choice. A third question then asked, without any restrictions, how many of the problems would they like to try.

The SAQ with three parts (see Appendix B) was administered to three classes of students representing a range in grade levels. In addition, individual interviews were conducted as students worked the SAQ. Two modes of responding were used: dichotomous (Agree/Disagree) and a five-point scale (Agree 1 2 3 4 5 Disagree). Appendix C contains tables comparing means between the two forms for each item in Part I. As a result of these tryouts a few rewording changes were made in Part I. More questions were raised about the effectiveness of Parts II and III. Responses in Part II varied very little. Because of this and the difficulty that some students experienced in reading the problems, Part II was dropped.

In Part III students selected the problems they would most like to work. The reasons for their choices varied. From the individual interviews, the more capable students would select problems on the basis of attributes related to the structure of the problem, such as "it was the easiest" or "I knew I could work it." The other students would select problems for superficial reasons such as "it was the first one" or "I liked the story." It appeared that knowing only the student's problem preference may be misleading. To know why the student selected a particular problem was also needed. At the end of the year the student may select the same problem for different reasons. The change in reasons could be important. Part III was modified to include a list of reasons that the student could use to indicate why he selected a particular problem (see Appendix D).

The revised Student Attitude Questionnaire contained two parts. Part I had fourteen statements for the student to respond "agree" or "disagree" (using the five-point scale). Part II asked the student to select from a set of three, the one problem he would most like to work on and to give the reason why he selected the problem. The SAQ in this form was reviewed by representatives of all three centers at the evaluation meeting of the Wagon Wheel Conference.

The evaluation working session participants agreed that it was still unclear exactly what Part II, problem preference (formerly Part III), would measure in its present form. For this reason, Part II was deleted from the SAQ. A point was made that if dimensions of problems were identified that appeared to be important to problem solving then these could be used to form contrasting pairs of problems. For example, a

problem with two conditions could be compared with a problem with several conditions. A problem with an associated picture could be compared to one without a picture. The student would be asked to select one of the pair of problems that he would like to solve. This could be used to measure the student's preferences for certain problem dimensions.

It was decided to increase Part I to include twenty items. With the elimination of Part II, more time could be devoted to trying to measure more meaningfully the three attributes of willingness, perseverance, and self-confidence. The present form of Part I was thought to be too impersonal in that students were asked to respond to statements made by other students. It was felt that our information would be more valid if the items were changed to true-false and the student asked if the statement represented how he felt.

The I.U. Center generated some more items. Willingness was broken down into three dimensions "gutting", cooperating, and liking. "Gutting" represented the willingness to try something without regard to its difficulty. Cooperating represented the willingness to go along with something--it is the thing to do. Liking represented the willingness to engage in something because it is fun--I like to do it. Two items, one positive and one negative, were included for each of these three dimensions of willingness.

Perseverance was separated into three dimensions: obtain right answer, not premature closure, and stick-to-itiveness. One kind of perseverance is sticking with the task until the <u>right</u> answer is found. Another is sticking with the task until just <u>any</u> answer is found, that is, not stopping work before some closure has been reached. The third

is stick-to-itiveness, working on a problem for a long time and not giving up right away. This last dimension is not dependent on getting any type of an answer. As with the willingness scale, two items (positive and negative) were selected for each of these three dimensions of perseverance.

Self-confidence was divided into three dimensions: belief to succeed, comparison to others, and "guts." Belief to succeed is the idea that "I am a good problem solver and will succeed most of the time." Comparison to others represents the person's self-confidence with respect to his/her peers--I am better or worse compared to other students. The guts of self-confidence relates to the difficulty of the situation--I can solve hard problems. Four items (two positive and two negative) were included under belief to succeed. Two items were included under each of the other two dimensions.

The final SAQ appears as Appendix E. Table 1 lists each item in the scale and the dimension of the scale that the item represents.

TABLE 1

Item Numbers for Each Dimension of Each Attitude Scale

Scale	Dimension	Positive Item	Negative Item
Willingness	Gutting	5	15
	Cooperating	3	14
	Liking	17	2
Perseverance	Right Answer	16	4
	Not Premature Closure	8	1
	Stick-to-itiveness	10	18
Self-confidence	Belief to Succeed	9, 20	12, 19
	Comparison	11	6
	Guts	13	7

Problem-Solving Survey

The draft of the problem-solving survey brought to the Wagon Wheel evaluation working session consisted of two parts. Part I was the NLSMA scale X023 which was given by NLSMA to the X population in the fall of the fourth grade. X023 contains five questions in a multiple choice format and was rated by the NLSMA staff to be at the analysis level. This scale appeared to be appropriate for the needs of MPSP for two reasons. First, two of the problems (4, 5) require processes similar to those which the project is focusing on. One problem (5) relates to using a table and the other (4) to finding a pattern. Second, the multiple-choice format gives students an opportunity to respond. Some concern was expressed about having problems so difficult that some students could not do any of the problems and would become very frustrated. In a multiple-choice test, a student has a choice and can always guess if he/she is not sure of the answers. Other issues that supported the use of the NLSMA scale were the ease of grading multiple choice items and the availability of data from the administration of the scale to a national population.

Part II of the survey (open-ended) consists of problems for the students to work, showing all of their work on paper. In the first draft of the survey each student is given two sets of three problems. From each set the student selects one problem to work. The rationale for giving the students a choice of problems was that they would feel better if they could select a problem that they could relate to instead of being forced to solve a given problem.

Prior to the Wagon Wheel Conference, the two parts of the problem-solving survey were tried out with ten students, some individually and some in groups of two. The problems were tried both as multiple-choice and as open-ended questions. Most of the students could do at least one problem on the NLSMA scale X023. For problems 2 and 3, the students had the idea and could find the answer to the open-ended question. However, when asked to find a number sentence which could and could not be used to solve the problems, they had trouble. In Part II, the better students liked to have a choice of problems and were able to handle the situation. Other students, as with Part II of the SAQ, selected problems for cosmetic reasons and, in some cases, did not read more than one problem.

The members of the evaluation working session approved the use of the NLSMA scale, X023, for Part I of the problem-solving survey. However, the members did not feel that giving the students a choice of problems to solve in Part II was the most efficient way. Instead, it was decided to construct four forms for each grade level. Each form was to have three problems which represented a range of difficulty levels. Each student, then, would work on all of the problems in the form she/he was given, showing her/his work and answers on the paper.

A set of approximately forty problems was compiled after the Wagon Wheel Conference representing three types of problems: textbook word problems that had one or two steps; problems (like those found in a textbook) that required more than two steps and that were not directly translatable into a number sentence; and the process-puzzle problems like those in the project's problem bank. The latter type of problem

was selected to insure that problems were included which could be solved by using different processes: guess & test, tables, listing, and searching for a pattern.

The I.U. staff selected problems from the set of forty to construct the four forms for each of the three grade levels. Each form contained three problems, one of each type. In some cases the same problem was used at different grade levels. The first problem was selected to ensure almost certain success. The instructions to the students indicated that they were to write down how they solved the problems. One of the grade 4 forms (4A) was given to one class to see if the instructions were clear and how the students would respond to the directions. The response was very favorable. Almost every student wrote down his computations and some idea about working the problem. Every student solved at least one problem and 21 out of 23 students solved the first problem correctly.

Part I (the NLSMA items) of the final problem solving survey appears as Appendix G, and Part II (the open-ended problems) appears as Appendix I.

REFERENCES

Covington, M. V. A Childhood Attitude Inventory for Problem Solving. Journal of Educational Measurement, 1966, 3, 234.

Lester, F. K., Jr. and Webb, N. L. "An Evaluation Design for the Mathematical Problem Solving Project: Rationale and Proposed Implementation." A paper prepared for the Research Workshop on Problem Solving in Mathematics Education, Center for the Study of Learning and Teaching of Georgia, The University of Georgia, July, 1975.

APPENDICES TO

TECHNICAL REPORT IV: SUMMATIVE EVALUATION

APPENDIX A: Evaluation of Standardized Tests

APPENDIX B: Student Attitude Questionnaire (First Draft)

APPENDIX C: Mean Scores for Dichotomous and Continuous Forms for Each Item in Part I of the Student Attitude Questionnaire (First Draft)

APPENDIX D: Student Attitude Questionnaire (Second Draft)

APPENDIX E: Student Attitude Questionnaire (Final Form)

APPENDIX F: SAQ Validation Report

APPENDIX G: Problem Solving Survey - Part I

APPENDIX H: Stanford Achievement Test - Intermediate Level I

APPENDIX I: Problem Solving Survey - Part II (12 Forms)

(Appendices Under Separate Cover)